Hidden Markov Models and Neural Networks for Fault Detection in Dynamic Systems
(Extended Summary)

Padhraic Smyth
Jet Propulsion Laboratory 238-420
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109
email: *pjs@galway.jpl.nasa.gov*
tel: (818) 354 3768, fax: (818) 354 6825.

1 Introduction

Continuous online monitoring of complex dynamic systems is common in applications as diverse as industrial plant operations, telecommunications networks, and biomedical health monitoring. For monitoring purposes, the exact nature of the system under observation is typically not relevant provided there exist some measurements or symptoms which provide diagnostic information regarding the underlying system state. This paper discusses how both pattern recognition (in the form of neural networks) and hidden Markov models (HMM's) can be used to automatically monitor online data for fault detection purposes. Monitoring for anomalies or faults poses some technical problems which are not normally encountered in typical HMM applications such as speech recognition. In particular, the ability to detect data from previously unseen classes and the use of prior knowledge in constructing the Markov model are both essential in applications of this nature. The paper describes recent progress on these and related topics (in the context of fault detection) and uses a real-world application at JPL to illustrate the ideas.

2 Pattern Recognition in Fault Monitoring

2.1 Background on Fault Detection for Dynamic Systems

For linear systems where the system dynamics and sensor measurement process can be completely modelled in an accurate manner, a variety of optimal control-theoretic methods for fault detection can be derived based on state estimation and statistical analysis of the residual error signals [1]. In practice, however, particularly for large complex systems, it is common that the system model may not be that accurate or reliable. A common technique in such a situation is to fit a dynamic model to the relationship between the measured input and output signals of the system - the model is often an autoregressive-exogenous (ARX) model. The parameters $\underline{\theta}$ of the (fixed order) model are estimated on-line in real-time using observed input/output data. Hence, fault detection can be carried out by observing changes in the values of the estimated parameter values (or model coefficients) relative to some reference set. This method has become known as the parameter method of fault detection [2, 3].

In a typical complex system there may be little or no prior knowledge as to how the parameters will change in relation to particular fault conditions. Assume that there exists a labelled database of training data consisting of observed data collected under both normal and fault conditions, i.e., a set of' pairs $\{\underline{\theta}, \omega_i\}$, where $\underline{\theta} \in R^d$ is a d-dimensional symptom or feature vector, and where $\omega_1, \ldots, \omega_m$ are state labels (typically ω_1 is the normal state and the others are fault states). Hence, one can use supervised learning or classification techniques to learn a model which estimates the posterior probabilities of the system state given the observed data, $p(\omega_i|\underline{\theta})$. Standard feed-forward neural networks can be quite useful in diagnostic applications of this nature. In earlier work we have described a particular application of this idea to fault diagnosis of an electro-mechanical antenna pointing system [4]. It was found that a single hidden layer neural network discriminant outperformed a parametric Gaussian classifier when used to classify estimated ARX coefficients. Unlike other non-parametric classification methods such as nearest neighbour classifiers, feedforward neural networks provide reasonable estimates of posterior class probabilities [5, 6]. This is a significant advantage in real-world applications where the network may be a component of an overall decision making system, e.g., part of a hidden Markov model as in Section 3.

2.2 Detection of Novel States

A standard assumption is that there are m known mutually exclusive and exhaustive states (or "classes") of the system, $\omega_1, , \omega_m$. The "mutually exclusive" assumption is reasonable in many applications where multiple simultaneous failure are highly unlikely. However, the exhaustive assumption is somewhat unrealistic. In particular, for fault detection in a complex system composed of hundreds of thousands of components, there are a myriad of possible fault conditions which might occur. The probability of occurrence of any single condition is very small, but nonetheless there is a significant probability that at least one of these conditions will occur over some finite time. As an example, for the antenna application described in Section 4, there are a few well-known faults (such as tachometer failures) which occur with regularity and can be assigned specific fault states in advance: however it is not practical to assign states to all the other minor faults which might occur.

Hence, the question must be asked as to whether or not a discriminant classifier trained to distinguish data from m states, can identify data from a different, or *novel* state. The answer lies in a simple application of Bayes' rule. If the classifier is a pure *discriminant*, i.e., it directly models the posterior probability $p(\omega_i|\underline{\theta})$, then it is implicitly relying on the assumption of exhaustivity and cannot in principle detect novel data. A good example of this type of classifier is a feedforward neural network using sigmoidal activation functions. Essentially, if one gives such a trained network, new data which is far away from the training data in the feature space, it will produce a near certain classification decision for one of the existing classes because of the semi-global nature of the sigmoid model [7].

On the other hand, a *generative* model directly models the data likelihood $p(\underline{\theta}|\omega_i)$ and then determines posterior class probabilities by application of Bayes' rule. Examples of generative classifiers include parametric models such as Gaussian classifiers and memory-based methods such as kernel density estimators and near neighbour models. Generative models are by nature well suited to online adaptation, in particular, adaptation of the structure of the model such as the inclusion of a new class -- conversely, discriminant models are by nature difficult to adapt online. However, there is a trade-off: because generative models typically are doing more modelling than just searching for a decision boundary, they can be less efficient (than discriminant methods) in their use of the data. For example, generative models typically scale poorly with input dimensionality for fixed training sample size - see Dawid [8] and Smyth [7] for further discussion.

An obvious idea in practice is use both a generative and discriminative classifier and add an "m +1tl I" state to the model to cover "all other possible states" not accounted for by the known m states. Hence, the posterior estimates of the generative classifier are conditioned on whether or not the data comes from one of the m known classes. Let the symbol $\omega_{\{1,\ldots,?),\}}$ denote the event that the true system state is one of the known states, and let $p(\omega_{m+1}|\underline{\theta})$ be the posterior probability that, the data is from an unknown state. Hence, one can estimate the true posterior probability of individual known states as

$$\sim(``ii!) = p_d(\omega_i|\underline{\theta}, \omega_{\{1,\ldots,m\}}) \times (1 - p(\omega_{m+1}|\underline{\theta})), \qquad 1 \leq i \leq m \tag{1}$$

where $p_d(\omega_i|\underline{\theta}, \omega_{\{1,\ldots,m\}})$ is the posterior probability estimate of state r' as provided by a discriminative model. $p(\underline{\theta}|\omega_{\{1,\ldots,n\,\iota\}})$ is provided directly by the generative model which typically can be a mixture of Gaussians or a kernel density estimate over all of the training data (ignoring class labels). The calculation of $p(\omega_{m+1}|\underline{\theta})$ can then be obtained 'i Bayes' rule 'f $p(\underline{\theta}|\omega_{m+1})$'s somehow known - in [9] an approach is described which uses a non-informative uniform Bayesian prior for $p(\underline{\theta}|\omega_{m+1})$ over a bounded space of parameter values.

3 Hidden Markov models for Temporal Context

3.1 General Principles

It is assumed that the reader is familiar with the basic terminology and concepts of HMM methods due to space limitations, detailed explanations and equations involving HMM's are omitted in this version of the paper (see Rabiner [10] for a thorough overview).

The methods outlined in Section 2 ignore temporal information in the sense that a given feature vector is classified without, using any information about previous features or classification estimates. Clearly this "instantaneous" classification] ignores the temporal context of the problem. For example, with the antenna application, the sampling interval between feature vectors is 4 seconds, while the mean time between failures is at least on the order of hours.

An elegant model which incorporates temporal context is that of' the discrete-time, finite-state, hidden Markov model. Taking a cue from the development of hybrid neural network/HMM applications in speech recognition, the idea is to embed the instantaneous estimate provided by the network within a Markov model framework [1 1]. Rather than directly modelling the correlations at the feature level, temporal correlation is directly modelled at the state level. The Markov model transition parameters can be estimated using a combination of prior knowledge of the long-term system behaviour and gross failure statistics (see Section 3.2). The two primary assumptions, namely, independence of feature estimates from one window to the next and a first-orcler Markov state dependence, both appear quite robust in practice provided certain reasonable assumptions are met [1 1]. The idea of treating the online monitoring of dynamic systems within the framework of a HMM appears to have been proposed independently by Smyth and Mellstrom [4] for electromechanical system monitoring and by Ayanoglu [1 ?] for communications networks modelling -- Provan [13] also describes a novel application of essentially the same idea to the problem of medical diagnosis and treatment of acute abdominal pain.

3.2 Specification of HMM Transition Probabilities

The nature of the II MM method used for online monitoring is considerably different to the more traditional HMM approach for speech recognition and language modelling. Ty pically, there is only a single model and the focus is 011 determining the likely state sequence for that particular model.

The model itself is dominated by a single state ω_1 describing normal system behaviour -- the system typically spends most of its time in this one state. Hence, unlike speech applications, the dynamics of the HMM can be quite simple - nonetheless, the practical benefits from modelling a system in this manner (as opposed to ignoring temporal context) can be significant in terms of improving the accuracy of the state estimates. In practice it is convenient to augment the model with states such as an "off" state where the system is no longer operating in its normal dynamic mode, e.g., the power is switched off to an electrical system.

As described in Section 2, the m -1 1th state ω_{m+1} is also unusual in the sense that it accounts for data from states which are unknown *a priori*. The incorporation of Equation (1) into the more standard HMM updating equations is straightforward and will not be described here.

While there may be training data for specific states available (normal data and fault data obtained under controlled conditions), it is unlikely that there are sequences of annotated data available for training. Hence, direct application Of Baum-Welch (or similar) methods for parameter estimation arc not feasible. Instead prior knowledge of the failure modes of the system must be used. For example, by use of the properties of the geometric distribution of state durations (for a first order HMM), the following simple relation holds between the transition probability a_{11} and the MTBF, t_{MTBF}, of the system:

$$a_{11} = 1 - \frac{\tau}{t_{MTBF}}$$

where τ is the sampling interval between state estimates. The MTBF can be either estimated from a problem database or from reliability specifications. Similarly, component failure rates for specific components provide constraints on other transition probabilities. In this manner, the HMM transition matrix is designed based on prior knowledge of overall system reliability [11]. In Section 5 methods for updating these estimates online are discussed.

Note that alternative approaches (to the HMM method) would be either to use "windows" of past data combined with present data as in time-delay neural networks, or to use recurrent neural networks with feedback. However, the HMM model has the distinct advantage of using prior knowledge in an elegant and explicit form, which in turn facilitates design and tuning of the monitoring model in an applications setting. For example, when aging components are replaced by new components, the HMM parameters Call be adjusted as a function Of lifetime reliability models.

4 Application to Antenna Monitoring

4.1 Problem Background

The system being monitored is a large steerable 34m ground antenna which is a critical part of NASA's Deep Space Network (DSN). There are 3 antenna sites (located in California, Spain and Australia) providing full 2A-hour coverage for deep space communications with various interplanetary spacecraft. The antennas represent critical potential single points of failure in the network. The antenna servo pointing systems are a complex mix of electro-mechanical components. A faulty component manifests itself indirectly via a change ill the characteristics of observed sensor readings in the pointing control loop. Because of the non-linearity and feedback present, direct causal relationships between fault conditions and observed symptoms can be difficult to establish - this makes manual fault diagnosis a time-consuming and expensive process and the application of direct analytical models impractical.

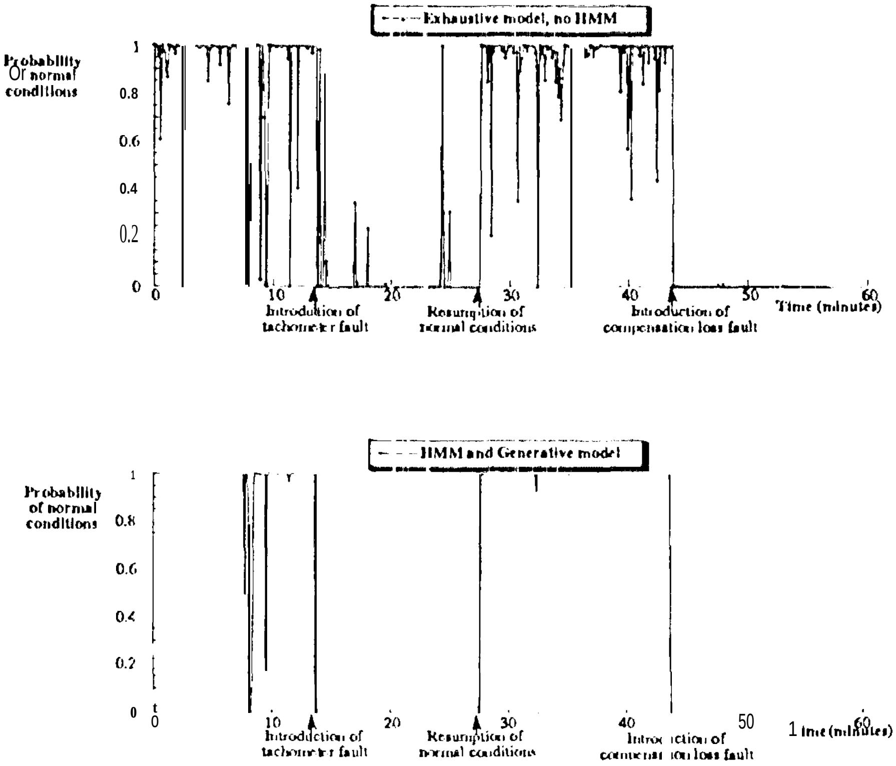

Figure 1: Estimated posterior probability of normal state (a) using no HMM and the exhaustive assumption (normal + 3 fault states), (b) using a HMM with a generative model (normal + 3 faults + other state).

4.2 Experimental Results

The neural- HMM method outlined in Sections 2 and 3 was applied to monitoring the motor current signal of the elevation pointing system for a DSN 34m antenna at Goldstone, California. Figure 1 (a) and (b) show typical comparative results for two different models (space limitations precluded the inclusion of more detailed experimental results). The models were trained and tested 011 independent data sets. The experiment consisted of introducing hardware faults into the system in a controlled manner at 15 minutes and 45 minutes, each of 15 minutes duration. Shown in Figure 1 are the model's estimates over time that the system is in the normal state -- corresponding graphs of estimates for the other states are not shown but are qualitatively similar. The input feature vector $\underline{\theta}$ consisted of 8 ARX coefficients and 4 energy estimates -- these were estimated from blocks of sensor data (motor current, tachometer, counter torque readings sampled at 50Hz) spaced at 4 second intervals.

Model (a) uses no HMM and assumes that the 4 known states are exhaustive -- a single feedforward neural network with 8 hidden units was used as the discriminative model. Model (b) uses a HMM with 5 states, where a generative model (a Gaussian mixture model) and a flat prior are used to determine the probability of the 5th state (as in [9]) and the same neural network as in model (a) is used as a discriminator for the other 4 known states (as in Equation (])). The parameters of the HMM were set according to known MTBF statistics for the system and the individual components.

The results indicate that model (a)'s estimates are quite noisy and contain a significant number of potential false alarms (highly undesirable in an operational environment). Model (b) is much more stable due to the smoothing effect of the HMM. Nonetheless, we note that between tile 8th and 10 minutes, there appear to be some possible false alarms. On inspection of the original data it was found that large transients (of unknown origin) were in fact present in the sensor data and that this was what the model had detected. The model without a generative component (either with or without the HMM) incorrectly classified this state as one of the known fault states (these results are not shown). A variety of other experiments have been carried out investigating the effects of different types Of density/discrimination models and Markov structures. The results have been quite robust. At present, the software implementation of model (b) is being tested for online daily use at one of JPL's antennas.

5 Work in Progress

An algorithm which could identify new transient states and add them to the model would be quite useful for autonomous operations. The problem of adapting HMM model structure is quite difficult however without the use of significant prior constraints - Stolcke and Omohundro [14] have developed a method for discrete HMM's, but the continuous density case is conceptually more complex and problematic.

The easier problem is that of adapting the parameters of a fixed model. Repeated use of the Baum-Welch estimation method in batch-mode seems inappropriate and computationally infeasible for online adaptation. Instead we have investigated simpler methods. In particular, for the transition probabilities, the use of Dirichlet priors [15] leads to intuitive and computationally simple updating schemes in addition, the method appears to be an approximation to the full Baum-Welch estimation process. We hope to have more results regarding online adaptation by the time of workshop.

Other issues involve the robustness of linear ARX models for characterising changes in a time series: currently we use a fixed order model where the order is chosen based on the normal data but may not be appropriate for fault or transient states. For the purposes of change detection the linear ARX features have worked well in practice -- however, less parametric representations would in principle provide the ability to detect a broader set of fault conditions.

Another primary limitation of the current model is the reliance on a first-order Markov assumption - again, extensions to include specific state duration density models (other than the implicit geometric density) for certain types of faults are under investigation - such semi-Markov processes have been successfully used in speech applications [16].

6 Conclusion

For fault monitoring applications the evidence so far indicates that the generative/discriminative approach is more sensitive to change detection than the purely discriminative method, and that the use of HMM methods increases the quality of the model substantially by using prior knowledge to account for temporal correlations. The ability to link these different models within a unifying probabilistic framework is a critical factor in their successful application.

Acknowledgements

The research described in this paper was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and was supported in part by D ARPA under grant number N00014- 92-J-1860

References

1. A. S. Willsky, 'A survey of design methods for failure detec tion in dynamic systems,' *Automatica*, pp.(ml--611,1976.

2. R. Isermann, 'Process fault detection based on modeling and estimation methods - a survey,' *Automatica*, vol.20, 387-404, 1984.

3. P. M. Frank, 'Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy a survey and somenew results,' *Automatica*, vol.26, no.3, pp.459- 474, 1990.

4. 1'. Smyth and J. Mellstrom, 'Fault diagnosis of antenna pointing systems using hybrid neural networks and signal processing techniques,' in *Advances in Neural Infor ma tion Processing Systems* **4,** R. Lippmann (cd.), Morgan Kaufmann Publishers: Los Altos, CA, 1992.

5. M. 1). Richard and R. P. Lippmann, 'Neural network classifiers estimate Bayesian a posteriori probabilities,' *Neural Computation*, 3(4), pp.461-483, 1992.

(i. J. Miller, R. Goodman, and 1'. Smyth, 'On loss functions which minimize to conditional expected values and posterior probabi lities,' to appear in *IEEE Transactions on Information Theory*.

7. 1'. Smyth, 'Probability density estimation and local basis function neural networks,' in *Computational Learning Theory and Natural Learning Syste ms*, T. Petsche, M. Kearns, S. Hanson, R. Rivest (eds.), Cambridge, MA: MIT Press, to appear, 1992.

8. A. 1'. Dawid, 'Properties of diagnostic data distributions,' *Biometrics*, 32, pp.647- 658, Sept. 1976.

9. P. Smyth and J. Mellstrom, 'Failure detection in dynamic systems: model construction without fault training data,' *Telecommuncations and Data Acquisition Progress Report 42- 112*, cd. E. C. Posner, Jet Propulsion Laboratory, Pasadena, CA, in p ress.

10. L. R. Rabiner, 'A tutorial o11 hidden Markov models and selected applications in s peech recognition,' *Proc. IEEE*, vol.77, no.2, pp.257- 286, Febr uary 1989.

11. 1'. Smyth, 'Hidden Markov models for fa ult detection in dynamic systems,' submitted for publication, 1992.

12. E. Ayanoglu, 'Robust and fast failure detection and prediction for fault tolerant communication networks," *Electronics Letters*, *28(10),* pp.940- *941,* 1992.

13. G. M. Provan, 'Modelling the dynamics of diagnosis and treatment using temporal influence diagrams,' in *Proceedings of the Third Workshop on Diagnosis*, pp.97- 106, 1992.

14. A. Stolcke aud S. Omohundro, 'Hidden Markov model induction by Bayesian merging,' to appear in *Advances in Neural Information Processing Systems 5*, C. L. Giits, S. J. Hanson and J. D. Cowan (eds.), San Make, CA, Morgan Kaufmann, 1993.

15. J. 0. Berger, *Statistical Decision Th eory and Bayesian* **Analysis,** 2nd cd., Springer Verlag: New York, 1985.

16. S. E. Levinson, 'Continuously variable duration hidden Markov models for automatic speech recognition,' *Computer*, Speech *amd Languag e*, vol.1, no.1, pp.29- 45, 1986.